A data augmentation approach to high dimensional ABC

Dennis Prangle

Newcastle University, UK

March 2016

Overview

Standard ABC works poorly with high dimensional data - a major drawback

This talk is preliminary work on an approach to deal with this

Joint work with Theodore Kypraios (Nottingham) and Richard Everitt (Reading)

Motivation

ABC background

Given:

- Observed data y_{obs}
- Probability model $\pi(y|\theta)$
  - Likelihood cannot be evaluated
  - Simulation from model straightforward
- Prior $\pi(\theta)$

Aim:

- Approximate the posterior $\pi(\theta|y_{\text{obs}})$

ABC rejection sampling

1. Sample θ from prior
2. Sample y from model
3. If $d(y, y_{\text{obs}}) \leq \epsilon$ accept
4. Return to step 1

Output: sample of θs from an approximate posterior

$$\propto \int \pi(\theta)\pi(y|\theta)\mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon]dy$$

$$= \pi(\theta)\tilde{L}_{\text{ABC}}(\theta)$$

ABC rejection sampling

1. Sample θ from prior
2. Sample y from model
3. If $d(y, y_{\text{obs}}) \leq \epsilon$ accept
4. Return to step 1

Output: sample of θs from an approximate posterior

$$\begin{aligned}&\propto \int \pi(\theta)\pi(y|\theta)\mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon]dy \\ &= \pi(\theta)\tilde{L}_{\text{ABC}}(\theta)\end{aligned}$$

Likelihood estimation interpretation

Can be viewed as importance sampling with a random likelihood:

$$\mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon]$$

i.e. estimate is 1 when y sufficiently close to y_{obs}
and zero otherwise

Target is the same as for the expectation of this:

$$\int \pi(y|\theta)\mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon] dy = \tilde{L}_{\text{ABC}}(\theta)$$

Likelihood estimation interpretation

Can be viewed as importance sampling with a random likelihood:

$$\mathbb{1}[d(y, y_{\mathrm{obs}}) \leq \epsilon]$$

i.e. estimate is 1 when y sufficiently close to y_{obs}

and zero otherwise

Target is the same as for the expectation of this:

$$\int \pi(y|\theta)\mathbb{1}[d(y, y_{\mathrm{obs}}) \leq \epsilon] dy = \tilde{L}_{\mathrm{ABC}}(\theta)$$

Motivation

ABC uses single sample rejection sampling estimate of $\tilde{L}_{\mathrm{ABC}}(\theta)$

Rejection sampling is poor when $\dim y$ is large: the probability of acceptance is very small

This project looks for a more efficient estimate.

Sketch of proposed approach

Input: a particular choice of θ:

- Draw several simulated datasets
- Perturb and refine the datasets in an attempt to improve their matches to y_{obs}
- Keep track of how likely all steps are

Output: an estimate of $\tilde{L}_{\text{ABC}}(\theta)$

This will be formalised as a SMC (sequential Monte Carlo) algorithm

Perturbations will be based on data augmentation ideas

(c.f. Andrieu et al 2012)

Sketch of proposed approach

Input: a particular choice of θ:

Draw several simulated datasets

Perturb and refine the datasets in an attempt to improve their matches to y_{obs}

Keep track of how likely all steps are

Output: an estimate of $\tilde{L}_{\text{ABC}}(\theta)$

This will be formalised as a SMC (sequential Monte Carlo) algorithm

Perturbations will be based on data augmentation ideas

(c.f. Andrieu et al 2012)

ABC curse of dimensionality

Intuition

Two sources of error in ABC are:

1. Poor target approximation of posterior: ϵ too high
2. Low acceptance rate: ϵ too low

Choice of ϵ involves a trade-off between these errors

As $\dim y$ increases error 2 becomes more problematic

And the optimal trade-off gets worse

Intuition

Two sources of error in ABC are:

1. Poor target approximation of posterior: ϵ too high
2. Low acceptance rate: ϵ too low

Choice of ϵ involves a trade-off between these errors

As $\dim y$ increases error 2 becomes more problematic

And the optimal trade-off gets worse

Results

MSE of ABC estimate under optimal tuning (i.e. ϵ etc) is

$$O_p(n^{-4/(4+\dim y)})$$

See Barber Voss and Webster (2015)

Above is for plain rejection sampling ABC

Similar results/heuristics for other ABC algorithms

Results

MSE of ABC estimate under optimal tuning (i.e. ϵ etc) is

$$O_p(n^{-4/(4+\dim y)})$$

See Barber Voss and Webster (2015)

Above is for plain rejection sampling ABC

Similar results/heuristics for other ABC algorithms

Summary statistics

Main strategy to avoid curse of dimensionality is dimension reduction

Replace high dimensional data y with lower dimensional summaries $s(y)$

i.e. accept if $s(y) \approx s(y_{\text{obs}})$ instead of $y \approx y_{\text{obs}}$

Reduces curse of dimensionality

But typically some information lost - another source of error

And we must choose which summaries to use

Ideally we'd like to avoid this difficult step

Summary statistics

Main strategy to avoid curse of dimensionality is dimension reduction

Replace high dimensional data y with lower dimensional summaries $s(y)$

i.e. accept if $s(y) \approx s(y_{\text{obs}})$ instead of $y \approx y_{\text{obs}}$

Reduces curse of dimensionality

But typically some information lost - another source of error

And we must choose which summaries to use

Ideally we'd like to avoid this difficult step

Summary statistics

Main strategy to avoid curse of dimensionality is dimension reduction

Replace high dimensional data y with lower dimensional summaries $s(y)$

i.e. accept if $s(y) \approx s(y_{\text{obs}})$ instead of $y \approx y_{\text{obs}}$

Reduces curse of dimensionality

But typically some information lost - another source of error

And we must choose which summaries to use

Ideally we'd like to avoid this difficult step

Other approaches to high dimensional ABC

- ABC-EP (Barthelmé et al)
- Sophisticated regression/classification (Pudlo et al)
- Using different summaries for each parameter in ABC MCMC (Wegmann et al)
- Combining marginal analyses (Nott et al)
- Neural network density estimation (Murray)

All involve some further approximations and/or costs

Other approaches to high dimensional ABC

- ABC-EP (Barthelmé et al)
- Sophisticated regression/classification (Pudlo et al)
- Using different summaries for each parameter in ABC MCMC (Wegmann et al)
- Combining marginal analyses (Nott et al)
- Neural network density estimation (Murray)

All involve some further approximations and/or costs

ABC likelihood approximation

Weighting kernel

Let $k_t(y)$ be weighting kernels

Each is a symmetric pdf with mode y_{obs}

And $\lim_{t\to\infty} k_t(y) = \delta_{y_{\text{obs}}}(y)$

e.g. Gaussian

$$k_t(y) \propto \exp\left[-\frac{d(y, y_{\text{obs}})^2}{2\epsilon_t^2}\right]$$

where $\epsilon_t \to 0$

or uniform

$$k_t(y) \propto \mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon_t]$$

Weighting kernel

Let $k_t(y)$ be weighting kernels

Each is a symmetric pdf with mode y_{obs}

And $\lim_{t\to\infty} k_t(y) = \delta_{y_{\text{obs}}}(y)$

e.g. Gaussian

$$k_t(y) \propto \exp\left[-\frac{d(y, y_{\text{obs}})^2}{2\epsilon_t^2}\right]$$

where $\epsilon_t \to 0$

or uniform

$$k_t(y) \propto \mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon_t]$$

Approximate likelihoods

Consider the approximate likelihood

$$L_{\mathrm{ABC},t}(\theta) = \int \pi(y|\theta) k_t(y) dy$$

Note that $\lim_{t\to\infty} L_{\mathrm{ABC},t}(\theta) = \pi(y_{\mathrm{obs}}|\theta)$, the true likelihood

Also, under a uniform kernel $L_{\mathrm{ABC},t}(\theta) \propto \tilde{L}_{\mathrm{ABC}}(\theta)$

Tempering scheme

Fix some value of θ

Define a sequence of unnormalised target densities

$$f_t(y) = \pi(y|\theta)k_t(y)$$

Let Z_t be the associated normalising constant i.e.

$$Z_t = \int \pi(y|\theta)k_t(y)dy$$

This equals $L_{\text{ABC},t}$

i.e. ABC likelihoods can be viewed as intractable normalising constants

Tempering scheme

Fix some value of θ

Define a sequence of unnormalised target densities

$$f_t(y) = \pi(y|\theta)k_t(y)$$

Let Z_t be the associated normalising constant i.e.

$$Z_t = \int \pi(y|\theta)k_t(y)dy$$

This equals $L_{\mathrm{ABC},t}$

i.e. ABC likelihoods can be viewed as intractable normalising constants

Ideal SMC scheme

Perform SMC with unnormalised targets $f_t(y)$

This let us form an unbiased estimate of Z_T/Z_1

Ensure $Z_1 = 1$

(e.g. Gaussian weight with $\epsilon_1 = \infty$ to give $f_t(x) = \pi(x|\theta)$)

We now have an unbiased estimate of the ABC likelihood $L_{\text{ABC},T}$

Problem: the $f_t(y)$s are intractable as they involve $\pi(y|\theta)$

Proposed solution: data augmentation

Ideal SMC scheme

Perform SMC with unnormalised targets $f_t(y)$

This let us form an unbiased estimate of Z_T/Z_1

Ensure $Z_1 = 1$

(e.g. Gaussian weight with $\epsilon_1 = \infty$ to give $f_t(x) = \pi(x|\theta)$)

We now have an unbiased estimate of the ABC likelihood $L_{\mathrm{ABC},T}$

Problem: the $f_t(y)$s are intractable as they involve $\pi(y|\theta)$

Proposed solution: data augmentation

Data augmentation approach

Model assumptions I

Suppose there are latent variables x

Such that $\pi(x, y|\theta)$ is tractable

and $y = y(x)$ (a deterministic function)

Can think of x as the full details of a simulation process

And $y(x)$ as partial observations

Then $\pi(x, y|\theta) = \pi(x|\theta)$

Model assumptions I

Suppose there are latent variables x

Such that $\pi(x, y|\theta)$ is tractable

and $y = y(x)$ (a deterministic function)

Can think of x as the full details of a simulation process

And $y(x)$ as partial observations

Then $\pi(x, y|\theta) = \pi(x|\theta)$

Model assumptions II

Assume that we have well behaved MCMC kernels targeting
$\pi(x|\theta)$

(Details of "well behaved" later)

Approximate likelihood

We can write our approximate likelihood in terms of x:

$$\begin{aligned} L_{\mathrm{ABC},t}(\theta) &= \int \pi(y|\theta)k_t(y)dy \\ &= \int \pi(x,y|\theta)k_t(y)dxdy \\ &= \int \pi(x|\theta)k_t(y(x))dx \end{aligned}$$

Tempering scheme

Fix some value of θ

Define a sequence of unnormalised target densities

$$f_t(x) = \pi(x|\theta)k_t(y(x))$$

Let Z_t be the associated normalising constant, then:

$$Z_t = \int \pi(x|\theta)k_t(y(x))dx$$

which equals $L_{\mathrm{ABC},t}$

SMC scheme

Perform SMC with unnormalised targets $f_t(x)$

This let us form an unbiased estimate of Z_T i.e. the ABC likelihood $L_{\text{ABC},T}$

As SMC forward kernel we use the data augmentation MCMC moves mentioned earlier

The kernel can be tuned at each step to aid mixing

SMC details

1. Set $t = 1$. Sample $x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(N)}$ from the model.

   Loop:

2. Increment t. Select new ϵ_t and Markov kernel K_t.
3. Update weights appropriately.
4. Terminate algorithm if ϵ_t equals a prespecified target.
5. If the effective sample size is below a prespecified threshold, resample the particles and update weights and likelihood estimate.
6. For $i = 1, \dots, N$ sample $x_t^{(i)} \sim K_t(x_{t-1}^{(i)})$.

   End loop

(c.f. Del Moral et al 2012)

Illustration: multivariate normal

Model

50 fixed locations $v_1, v_2, \ldots, v_{50}$ in $[0, 1]$

Model: $y_1, \ldots, y_{50} \sim N(0, \Sigma)$

Covariance function is

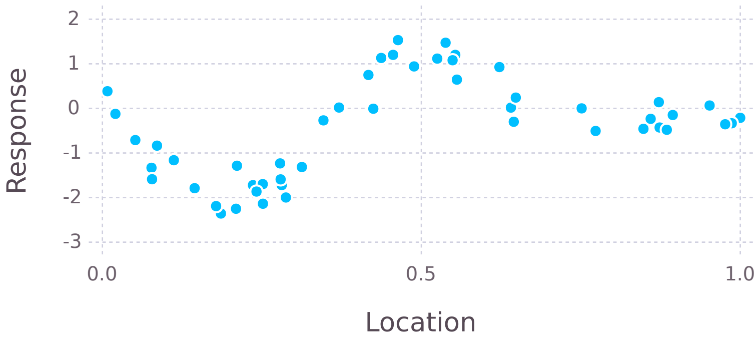
$$\rho(v, v') = 4 \exp(-[\tfrac{v-v'}{\phi}]^2) + 0.1 \mathbb{1}(v = v')$$

i.e. a squared exponential covariance function with variance 4 and scale ϕ plus a nugget effect

Inference for $\dim(y) = 50$ not feasible by standard ABC

Observed data

Pseudo-observations sampled from model with $\phi = 0.3$



Simulation results

200 particles
Each estimate took roughly 1 second ($\epsilon = 3.2$) to 10 seconds ($\epsilon = 0.1$)
Results improve as ϵ reduced

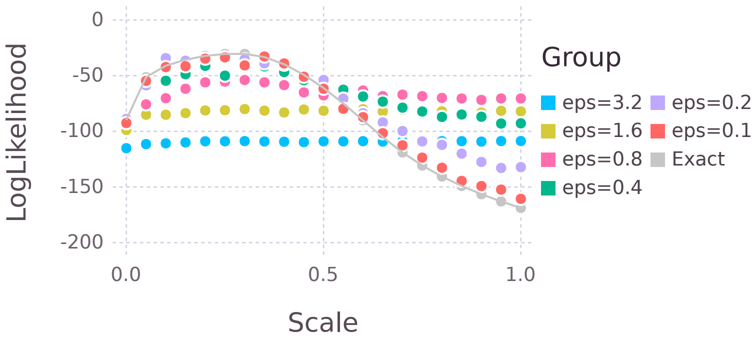


Illustration: SIR model

Model

Standard susceptible infectious removed model

Homogeneous mixing, Markovian events

Removal times observed

2 parameters: infection and removal

Synthetic data

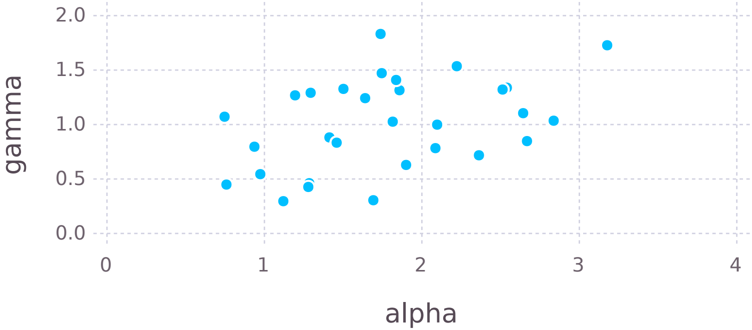
We can take x as some independent random variables

And observations $y(x)$ involve simulation by the Selke construction

Inference

I used Bayesian optimisation to get a rough posterior approximation

Then importance sampling to get more accurate results



Discussion

Summary

- Method proposed for estimation of intractable likelihoods
- Based on SMC rather than rejection sampling (as in ABC)
- Learns good simulations instead of randomly sampling them
- Uses full data instead of summaries
- Reasonable preliminary results for two simple examples

Limitations

- Need suitable MCMC moves for data augmentation scheme
  i.e. must be able to explore $\pi(x|\theta, y \approx y_{\text{obs}})$ easily

  Seems hard to achieve in some applications e.g. coalescent

- Also the overall method can be very expensive

Future work

- Intractable SIR model: missing/censored data
- Best way to use likelihood estimates in an inference method

  e.g. SMC^2?
- Reduce computational cost

  e.g. via particle Gibbs, auxiliary variable methods, delayed acceptance
- Theory

  How does complexity scale?
  Characterise when more efficient than ABC